Hypothesis testing complexity in the name of ethics: Response

Knottnerus, J.A.; Bouter, L.M.

published in
Journal of Clinical Epidemiology
2002

DOI (link to publisher)
10.1016/S0895-4356(01)00459-0

document version
Publisher's PDF, also known as Version of record

Response

Moyé and Tita basically address two issues. First, in their view, one-sided testing should be avoided in health care research, especially in randomized trials, and is typically motivated by the investigators' *a priori* intuition or suggestive information from prior studies, or "belief." Second, they argue that "the two-sided test shines bright, direct light on the possible production of harm by the intervention," especially unsuspected harm. On both issues we do not agree with their views.

Both the one-sided and two-sided approach presuppose a good balance between uncertainty (otherwise the trial would not be necessary) and plausibility (otherwise the trial would not be justified) of a relevant hypothesis based on clinical and scientific insights. One-sided testing is appropriate and even indicated if such testing is sufficient to answer the central research question. This is the case if (1) the scientific hypothesis to be tested is obviously one sided, or if (2) only a clear advantage in effect of the principal over the reference intervention would have consequences for practice, for instance, if the principal intervention implies a more burdensome regimen for the patient. An example of (1) would be an exploratory trial comparing a new intervention with a placebo. Only a clear benefit for the new intervention will make further research indicated. No difference, or superiority of the placebo, would lead to the end of research on this new intervention. This "one-sided" scientific hypothesis of a possible relevant effect, not "a strong belief in the benefit of a therapy," is to be confirmed or rejected. Such an hypothesis is relevant irrespective of individual beliefs, although a very strong (dis)belief of investigators would be a contraindication to perform the trial. Regarding (2), we cannot understand why investigators should prove that a principal intervention is worse than a reference intervention, instead of "not better," if applying the principal intervention in clinical practice would only be acceptable if it were clearly more effective.

We emphasize again, that the price of not sharing our view is high in terms of subjects to be included and clinical end points needed. Elaborating our numerical examples in the original Commentary, we get the picture of Table 1. Looking at the situation with the higher incidence (50% in the reference group) in the two-sided option, not only 166 more subjects must be included in the trial, but the expected total number of "end points" would be 349 before it could be decided that the principal treatment is better, not better, or worse as to the primary end points. In the one-sided option, the expected number of end points would be 275 before it could be decided that the principal treatment is better or not better. Especially if the end points would be severe morbidity or death, which is often the case, the difference of 74 is huge. In the example with the lower incidence (10% in the reference group), the corresponding differences are smaller but still relevant. So, the price for two-sided testing is so high that it would be only acceptable if a convincing reason is given. In other words, we advocate one-sided testing as the default option: "one-sided, unless . . ." instead of "two sided, unless . . ."

The sample size of an effectiveness trial is neither explicitly estimated nor generally appropriate to detect unsuspected harm. We already argued that for both practical and scientific purposes it is often not necessary to establish the statistical significance of inferiority of the principal intervention. As far as unsuspected adverse events are concerned, these can typically not be taken as the basis for sample size estimation. Moreover, trials will almost never have sufficient power to detect very low rates of severe adverse effects that could be clinically relevant. Detecting unsuspected adverse effects would require enormous sample sizes that will still never be large enough to detect them all. In addition, the approach advocated by Moyé and Tita might er-

Table 1
Required numbers per comparison group and expected numbers of primary end points (smoothed to absolute numbers) to detect a minimal cumulative incidence reduction of 10%, starting from one- and two-sided testing, respectively, for expected cumulative incidences of 20% and 50% in the reference group (equal group sizes, type I error = 0.05, type II error = 0.20).

Minimal incidence reduction to be detected	Group size and number of endpoints	One-sided testing	Two-sided testing
reference group: 20%	Required group size	157	199
principal group: 10%	Primary endpoints		
	reference group	31	40
	principal group	16	20
	total	47	60
	difference reference-principal	15	20
reference group: 50%	Required group size	305	388
principal group: 40%	Primary endpoints		
	reference group	153	194
	principal group	122	155
	total	275	349
	difference reference-principal	31	39

Source: Knottnerus JA, Bouter LM, Ned Tijdschr Geneeskd 2001;145:1051–4.

roneously suggest that the principal intervention is safe when, in the context of an effectiveness study, a two-sided approach has not demonstrated clinically relevant adverse effects. However, one should not mix up studies aimed to study effectiveness as to primary clinical end points (which may also include adverse effects to be intentionally evaluated, for example when looking at the incidence of both stroke and bleeding in a warfarin trial) with studies especially designed to detect or exclude infrequent but severe, and unsuspected adverse effects. The latter type of studies, either experimental or observational, have specific requirements, regarding both methodology and sample size estimation [1]. Of course, it is wise to use infrequent adverse effects earlier observed in effectiveness trials in preparing such studies. Also, we must always be keen to take action after observing unsuspected frequent adverse effects in effectiveness trials. For such situations an independent and thoughtful case-by-case judgement on the necessity of immediate action or further evaluation is important, rather than the idea that a default two-sided approach would have avoided further study.

If a one-sided trial shows no benefit (but possible harm), repetition is not necessary after such a trial, because the study question (in this case, whether the principal treatment would be better) has already been answered. Trials should be focused on specific questions. It would be ethically disputable to collect more events than needed for answering the specific question participants were invited for, with the motivation that detection of harm among these participants might be useful in view of additional, possibly future, indications. Other indications should be studied as such, with a tailor-made specific design.

Indeed, if feasible, stopping rules can prevent unnecessary (and thus unethical) allocation to inferior interventions. But the possibility to apply stopping rules is not at all the monopoly of the two-sided approach. Such rules can also be implemented according to a one-sided approach both regarding primary endpoints and adverse effects. [2–4]. In general, however, we must keep in mind that the feasibility of stopping rules depends on a combination of slow recruitment and a short follow-up to the primary end point.

For us, “complexity” of two-sided testing nor “simplicity” of one-sided testing, nor any relation between complexity and ethics, is at stake in choosing between a one- or two-sided approach. We just say: in many effectiveness trials, to choose a one-sided approach is appropriate. In their study design, researchers should explicitly describe why the one- or two-sided approach is their starting point for sample size estimation. If an investigator thinks that more subjects should be included or more “end points” are needed before a conclusion in a clinical trial can be reached, in the context of a two-sided instead of a one-sided approach, this should be positively motivated and critically reviewed in advance.

In summary, we believe that the one-sided approach should be the default option.

J. André Knottnerus
Netherlands School of Primary Care Research
University of Maastricht
Maastricht, The Netherlands
Lex M. Bouter
Institute for Research in Extramural Medicine
(EMGO Institute)
Amsterdam, The Netherlands

References

[1] Turbert-Bitter P, Manfredi R, Lellouch J, Begaud B. Sample size calculations for risk equivalence testing in pharmacoepidemiology. J Clin Epidemiol 2000;53:1268–74.
[2] Demets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. Biometrika 1980;67:651–60.
[3] Berry DA, Ho CH. One-side stopping boundaries for clinical trials: decision–theoretic approach. Biometrics 1988;44:219–27.
[4] Kumar KV, Powell MR, Waligora JM. Early stopping of aerospace medical trials: application of sequential principles. J Clin Pharmacol 1994;34:596–8.

PII: S0895-4356(01)00459-0